DC3 Digital Forensics Challenge

303 – Language Text Identification

Team Information

Team Name: elsø-ren

Results Email: [redacted]

Examination Time Frame: to

Instructions

Description: Examiners must develop and document a methodology used to identify the language presented in the documents in the **303_Language_Text_Identification_Challenge2008**.

Report the detailed explanation of your process (software or technique) used to determine the language and document information.

Total Weighted Points: 60 Total Points available per entry – Total 300 Points Available

1. **Answers** – Fill in the chart below with your findings. *As a Forensic Challenge, consider that your answers will have to have enough detail for the Findings and Methodology of your examination to satisfy questioning in a court of law.*
2. **Methodology** – Provide a meticulously detailed explanation of your process. Be sure to include a step action that our reviewers can follow to reproduce your work for authenticity including tools and techniques.

Internal Reviewer Use Only

Reviewer:

Points Awarded:

Date:

Review Period: to

Completed: □ Yes □ No □ Partial

Question 303: Language Text identification

Tools Available:

TextCat http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

Xerox: http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html

Lextek Language Identifier http://www.lextek.com/langid/li/onlineidentifier.html

Tools we used:

TextCat http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

Xerox: http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html

Process:

1) Using a webbrower go to the following website hosting the language ID tool:
http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

2) Baseline the language tool.

Verify that the Language identifying tool works with known data: go to
http://www.omniglot.com/babel/index.htm

Select text for about 4 languages and populate into the tool. Verify the answers are correct.

3) Open the language test data set in MS Word.

Copy the data and paste into the Language ID tool site.

Click on "Guess"

If language is identified, it will be displayed at the bottom.

Repeat step 3 for all other language texts.

4) Verify results using another tool:

Open site: Xerox: http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser.en.html

Open a language document, copy and paste the data into the tool

Click "Guess Language"

Compare results to what was obtained in step 3.

Repeat step 4 for all other data sets. Verify results, they must be the same.

Our results:

Document 5tX: Japanese

Document 4tx: Arabic

Document 3Tx: Korean

Document 2Tx: Russian

Document 1Tx: Chinese